Significance testing and forward stepwise with the group lasso

Jonathan Taylor* **and Joshua Loftus and ???**

Department of Statistics
Stanford University
Sequoia Hall
390 Serra Mall
Stanford, CA 94305, U.S.A.
e-mail: `jonathan.taylor@stanford.edu`

Abstract: We extend the work of Lockhart et al. (2013) on significance testing for the lasso to the group lasso. We derive the corresponding test statistic and its exact distribution, and show that these can be used with (non-penalized) forward stepwise regression and model selection. For group lasso the forward stepwise procedure adds a group of variables in each step. These procedures are shown to perform well in simulations and in a real data example.

AMS 2000 subject classifications: Primary 62M40; secondary 62H35.
Keywords and phrases: penalized regression, convex analysis, least squares, Gaussian processes.

1. Introduction

Forward stepwise regression is a stochastic model selection procedure that begins with an empty model and adds the best predictor variable in each step. Classical significance tests fail when a model has been selected this way and tend to be anti-conservative. Recently, Lockhart et al. (2013) found a novel test statistic with an appropriate null distribution that behaves well when a model has been selected using the lasso (Tibshirani, 1996). Taylor et al. (2013) modified and extended those results to the group lasso (Ming and Lin, 2005) and other adaptive regression problems. The present work explores the behavior of those test statistics for models selected by forward stepwise procedures and works out some of the details involved in applying these methods to models with grouped variables. We find that our test statistics can be used for valid significance testing, or for tempering the aggressive nature of forward stepwise model selection. The resulting methods may be less computationally intensive than choosing a regularization parameter for the group lasso by cross-validation.

To fix notation for the group lasso problem we consider an outcome $y \in \mathbb{R}^n$. Let an integer $G \geq 1$ be the number of groups (or factors). Lowercase g will be used as an index for groups. For each $1 \leq g \leq G$ the design matrix encoding the gth group is the $n \times p_g$ matrix denoted X_g. Throughout most of this paper we assume orthogonality within groups, that is $X_g^T X_g = I_{p_g \times p_g}$ for all g. (**To do:**

*Supported in part by NSF grant DMS 1208857 and AFOSR grant 113039.

Take this out if we don't actually need it). Let $p = \sum_{g=1}^{G} p_g$ be the total number of individual variables, and let X be the matrix constructed by column-binding the X_g, that is

$$X = \begin{pmatrix} X_1 & X_2 & \cdots & X_G \end{pmatrix}$$

With each group we associate the $p_g \times 1$ coefficient vector β_g, and write β for the $p \times 1$ vector constructed by stacking all of the β_g in order. Finally, our model for the data is

$$\begin{aligned} y &= X\beta + \sigma\epsilon \\ &= \sum_{g=1}^{G} X_g \beta_g + \sigma\epsilon \end{aligned} \tag{1}$$

where ϵ is noise. Unless otherwise specified we will assume i.i.d. Gaussian noise $\epsilon \sim N(0, I_{n\times n})$ throughout. So far we have only introduced some new notation to the usual linear regression setup, but now we define the *group lasso estimator* as the following solution to a convex problem

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} w_g \|\beta_g\|_2 \tag{2}$$

The parameter $\lambda \geq 0$ enforces sparsity in groups: for large λ most of the β_g will be zero vectors. The weights w_g are usually taken to be $\sqrt{p_g}$ to normalize the penalty across groups. Note that this includes the usual lasso estimator as a special case when all of the groups are of size 1, since then the penalty term is the L^1-norm of β.

The group lasso estimator is discussed in (ref that guy's thesis) and (ref Yuan and Lin). An important extension is the sparse group lasso (ref ???) which enforces sparsity in groups as well as sparsity of the coefficients within the groups. For a survey on group lasso and related factor models see (ref???). **To do:** review some more literature and add a few more references here if they seem worthwhile.

In a recent work Lockhart et al. (2013) defined a *covariance test statistic* for testing the significance of a variable entering the model along the lasso solution path. They derived a simple asymptotic null distribution, proved a type of convergence under broad "minimum growth" conditions, and demonstrated in simulations that the test statistic closely matches its asymptotic distribution even in finite samples. That work marked an important advance in the problem of combining inference with model selection. The current paper extends some of their results to the group lasso (ref Yuan and Lin?). In the process we had to derive an exact finite sample null distribution for the test statistic (ref TLTT?). We also show that the techniques used to get these results can be used to do a forward stepwise procedure related to the group lasso that adds groups of variables (or factors) in each step.

(**To do:** Change this paragraph as the sections are completed) In Section 2 we briefly review the work of (ref LTTT) concerning significance testing for the lasso, and discuss how this result is extended to group lasso. Section 2 also introduces more of the terminology we use to describe the present problem. Section 3 describes a framework for forward stepwise procedures using the test statistic derived in Section 2. In this framework groups of variables are added in each step until some stopping criterion involving the test statistic is satisfied. In Section 4 we show the results of some simulations for several forward stepwise procedures, and in Section 5 we apply our method to a real data set (**To do:** brief description of data).

2. Significance testing: from lasso to group lasso

2.1. Background

In the ordinary least squares setting, a significance test for a single variable can be conducted by comparing the drop in residual sums of squares (RSS) to a χ^2_1 distribution. Similarly, when adding a group of k variables we can compare the drop in RSS to a χ^2_k random variable. This generally does not work when the variable to be added has been chosen by a method that uses the data, and in particular it fails for forward stepwise procedures which add the "best" (e.g. most highly correlated) predictor in each step. In that case, the test statistic (drop in RSS) does not match the theoretical null distribution even when the null hypothesis is true. Lockhart et al. (ref LTTT) introduced a new test statistic based on the knots in the lasso solution path. Writing $\hat{\beta}(\lambda)$, the solution for a fixed value of λ, we need the following facts (ref Tibs2012)

- The vector valued function $\hat{\beta}(\lambda)$ is a continuous function of λ. For the lasso path, the coordinates of $\hat{\beta}(\lambda)$ are piecewise linear with changes in slope occurring at a finite number of λ values referred to as *knots*. The knots depend on the data and are usually written in order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$. We follow this convention.
- The *active set* A_k is a set of indices of variables for which the corresponding coordinates of $\hat{\beta}(\lambda_k)$ are potentially nonzero. Any variable with index not in A_k has a zero coefficient in $\hat{\beta}(\lambda_k)$, but the converse is not true.
- Path algorithms for computing lasso solutions proceed by fitting models at a grid of λ values. The active set changes whenever λ crosses a knot, and predictor variables can both enter and leave the active set. However, at the first two knots λ_1 and λ_2 no variable can leave the active set. So the first two knots correspond to the first two variables entering the model.

Lockhart et al. prove that, under the null hypothesis that A_k contains all the strong predictor variables, the distribution of a test statistic $T_k \propto \lambda_k(\lambda_k - \lambda_{k+1})$ is asymptotically Exp(1). In the lasso case we know a lot about the knots and active set, but the group lasso picture is slightly more complicated. For the group lasso, $\hat{\beta}(\lambda)$ does not have piecewise linear components. To overcome this difficulty we will restrict our attention to the first group of variables to enter the active set since the analysis then follows almost exactly as for the lasso. (**To do:** Change this if I can find λ_k).

Before considering the group lasso, we review some ingredients of the proof for the lasso. Let $J = \{1, 2, \ldots, p\}$ index variables and consider a stochastic process $f_{j,s} = sX_j^T y$ defined on $T = J \times [-1, 1]$. This stochastic process is simply a collection of linear combinations of y, hence it is Gaussian under the assumption of Gaussian errors. The *Karush-Kuhn-Tuckher (KKT) conditions* (ref???) imply that $\lambda_1 = \max_j |X_j^T y|$. By introducing the sign variable s, we can remove the absolute value and write λ_1 as the maximum of our Gaussian process

$$\lambda_1 = \max_{(j,s)} f_{j,s} \tag{3}$$

We have exhibited the first knot as the maximum of a Gaussian process. We can do this for the second knot by introducing a new process. Let (j_1, s_1) be the maximizer so that $\lambda_1 = s_1 X_{j_1}^T y$, and define

$$\begin{aligned} f_{(j,s)}^{(j_1,s_1)} &= \frac{sX_j^T y - sX_j^T X_{j_1} X_{j_1}^T y}{1 - ss_1 X_j^T X_{j_1}} \\ &= \frac{sX_j^T (I - P_{j_1}) y}{1 - s_1 X_{j_1}^T X_j s} \end{aligned} \tag{4}$$

where P_j is the projection onto the subspace spanned by X_j. We can think of this as a "residual process" after regressing out the maximum. Write $M = \max_{j \neq j_1, s} f_{(j,s)}^{(j_1,s_1)}$, the maximum of this residual process. It can be shown from the KKT conditions that $M = \lambda_2$. To summarize, we have represented the first two knots of the lasso solution path as the maxima of some natural Gaussian processes. Distributional facts about Gaussian processes now allow us to make conclusions about the distribution of functions of the knots.

2.2. Group lasso

To extend this argument to the group lasso we need to define Gaussian processes that characterize the knots of the group lasso solution path.

To do: Either use the simplified argument in the case of equal weights (equal group sizes), or finish adjusting the proof of $M = \lambda_2$ to include the weights and use that version. The proof can go in an appendix.

(For this section I have a lot of stuff written that can just be pasted in once I'm done wrangling the KKT conditions for the unequal weight case)

2.3. Better p-values

- **To do:** Add the figures to this section
- **To do:** Convert to exposition instead of bulleted list
- As in LTTT, $T = \lambda_1(\lambda_1 - M)$ and $M = \lambda_2$

- Convergence to the limiting Exp(1) distribution is too slow

$$\frac{P(\chi_k/w_g \geq m + t/m)}{P(\chi_k/w_g \geq m)} \to e^{-t} \text{ as } m \to \infty$$

  (when the group achieving λ_1 is group g and has rank k)
- The limiting distribution only depends on T, but we also observe M
- Let's just try the ratio (conditional χ_k tail probability) evaluated at T and M (it works better)
- Going one step further, instead of using the approximation (see LTTT Proof of Lemma 5)

$$\frac{M + \sqrt{M^2 + 4t}}{2} \approx M + \frac{t}{M}$$

  we can just use the left hand side
- For $T = \lambda_1(\lambda_1 - M)$ the left hand side simplifies to λ_1
- Now our p-value is

$$\frac{P(\chi_k/w_g \geq \lambda_1)}{P(\chi_k/w_g \geq \lambda_2)}$$

2.4. Can we do λ_k? Don't know yet

3. Forward stepwise procedures

To do: This is potentially the second largest thing not done yet. However, if it turns out we can't think of a better stopping rule, then this section is done and just needs to be written (and would not take long).

4. Simulations

(I've done many of these and have a lot of code, so I can produce a lot of output for this section) **To do:** Modify some of my previous simulations after deciding on a stopping rule, decide which simulations to put here.

5. Real data example

To do: This is definitely the single largest part I have not done yet.

References

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2013). A significance test for the lasso. *arXiv:1301.7161*. Submitted to Annals of Statistics, URL `http://arxiv.org/abs/1301.7161`.

Ming, Y. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, **68** 49–67.

Taylor, J., Loftus, J. and Tibshirani, R. (2013). Tests in adaptive regression via the kac-rice formula. *ArXiv e-prints*. `1308.3020`.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, **58** 267–288.